DOI: https://doi.org/10.15407/bv2023.03.003
УДК 025.45:004

Олександр Кузнєцов,
https://orcid.org/0000-0002-9902-1295,
завідувач відділу наукового комплектування та опрацювання бібліотечних фондів,
Державна науково-технічна бібліотека України (Київ, Україна)
вул. Антоновича 180, Київ, 03150, Україна
e-mail: nkof@dntb.gov.ua

Віктор Заїка,
https://orcid.org/0009-0003-6582-6524,
кандидат фізико-математичних наук,
провідний інженер відділу бібліометрії і наукометрії,
Національна бібліотека України імені В. І. Вернадського (Київ, Україна)
просп. Голосіївський, 3, Київ, 03039, Україна
e-mail: victor.zayika@gmail.com

Визначення індексів УДК нових надходжень в електронному вигляді для формування електронної бібліотеки програмними засобами

Мета статті – запропонувати методику аналізу достовірності індексу УДК повнотекстових документів, що надходять до бібліотеки від різних організацій і авторів в електронному форматі, продемонструвати її застосування на прикладі п’яти електронних документів економічної тематики (індекс УДК 331), використовуючи створений програмний інструмент «Аналіз текстів». **Методологія дослідження.** Застосовано кількісний метод дослідження змісту документів. Для знаходження подібних за змістом документів (файлів) використано поняття «косинусної міри подібності» та розраховано коефіцієнти тематичного напряму для кожного документа. Текстові файли представлено у вигляді векторів у багатомірному просторі. З цією метою різні словоформи було зведено до однієї лексеми та пораховано кількість (або частоту) вживання лексем у кожному документі. Лексеми протлумачено як координати, а частоту вживань – як значення відповідної координати. Після векторизації текстів застосовано математичний апарат аналітичної геометрії, а тематиці кожного текстового документа співставлено відповідне числове значення – коефіцієнт тематичного напряму. **Наукова новизна**. Вперше використано методи контент-аналізу, а саме кількісний аналіз, для оцінки достовірності індексу УДК документа, створено програмний інструмент, використання якого допоможе систематизатору підтвердити чи спростувати індекс УДК сумнівного документа не читаючи його. **Висновки.** Авторський програмний інструмент та запропонована методика корекції УДК можуть бути використані при створенні репозитаріїв електронних текстів, вони сприятимуть підвищенню якості інформаційного пошуку та вибору контенту. При накопиченні певної кількості електронних документів, завдяки розробленій методиці, УДК нового тексту (надходження) можна визначити автоматично за показником коефіцієнтів тематичного напряму (близько одиниці) нового тексту та відповідного корпусу. Вектор коефіцієнтів тематичного напряму текстів, що досліджувались, їх розподіл за зростанням коефіцієнтів тематичного напряму, дав змогу виявити кластер – групу однакових текстів за змістом. Достовірним критерієм є величина коефіцієнту при змінній лінійної апроксимації, в ідеалі горизонтальна поличка на графіку розподілу коефіцієнтів тематичного напряму – коефіцієнт дорівнює одиниці. Кількість тематичних напрямів визначається кількістю кластерів.

К л ю ч о в і с л о в а: системи комп’ютерного аналізу тексту, контент-аналіз, косинусна міра подібності, індекс УДК, кластер, електронна бібліотека, частотний масив, коефіцієнт тематичного напряму, програмні пакети для контент-аналізу.

Актуальність теми дослідження. Ми живемо в епоху стрімкого розвитку інформаційних технологій та лавиноподібного зростання обсягів різноманітної інформації, так званого інформаційного вибуху. Існують дослідження, в яких стверджується, що в найближчому майбутньому обсяг інформації подвоюватиметься кожні два роки [13]. Однією з причин стрімкого зростання інформації є збільшення частки даних, що генеруються автоматично, зокрема системами штучного інтелекту, такими як Chat GPT [3], в основі яких лежать сучасні лінгвістичні технології та доступ до енциклопедичних знань у вигляді баз даних. Chat GPT здатний підтримувати діалог ніби як людина, за запитом знаходити потрібну інформацію. Він може навіть створити наукову публікацію! Цей інструмент, з одного боку, допомагає продукувати інформацію, а з іншого – ускладнює пошук, оскільки створює надлишкову інформацію, яка часто не містить жодної новизни.

Частина інформації, яку генерує людство, є надзвичайно важливою. Щоб мати змогу знаходити потрібні важливі дані та відомості, їх необхідно класифікувати, систематизувати та зберігати належним чином. В Україні в бібліотеках, видавництвах, інформаційних центрах для систематизації документів, пошуку інформації за галузями знань використовуються таблиці УДК. УДК – універсальна десяткова класифікація – міжнародна багатомовна класифікаційна система, що об'єднує всі галузі знань в єдиній універсальній структурі з загальною десятковою ієрархією. Індекс УДК – основа для впорядкування накопичених людством знань у традиційних бібліотеках, електронних базах даних та інших сховищах інформації. Головна функція, яку виконують таблиці УДК, – максимально точно відобразити зміст видання і забезпечити в подальшому швидкий та легкий пошук інформації. Індекси УДК побудовані так, що кожна наступна цифра, яка приєднується до індексу, не змінює попереднє значення, а лише уточнює його. За допомогою класифікатора УДК легко знайти будь-яку інформацію із галузі мистецтва, літератури, науки. Загалом у світі понад 130 країн використовують УДК.

Індекс УДК дає змогу отримати уявлення щодо тематики, виду, типу літератури, не читаючи її. Присвоєння некоректного індексу УДК літературній одиниці може зробити її недосяжною для користувачів, при тому, що ця інформація представлена в сховищі та поглинає ресурси, необхідні для її зберігання.

Традиційно УДК присвоюється людиною, систематизатором, що аналітико-синтетично, тобто інтелектуально, опрацьовує літературне джерело, виокремлює його змістовні та формальні ознаки для того, щоб подати стиснену інформацію про цей документ в інформаційно-пошуковій системі у вигляді класифікаційних індексів. Хоча робота систематизатора значною мірою формалізована, оскільки він керується усталеним набором правил і таблиць, індекс УДК також є відображенням світогляду систематизатора, щоправда крізь призму вищезгаданих таблиць. З одного боку, послідовність дій систематизатора визначено заздалегідь, з іншого – даний процес містить творчу складову, тому люди з різним світоглядом та життєвим досвідом можуть по-різному виокремлювати змістовні та формальні ознаки одного й того ж документа. Крім того, щоб коректно сформулю-

вати змістовні та формальні ознаки (визначити індекс УДК) систематизатор має детально ознайомитися зі змістом, тобто контентом літературної одиниці, що на практиці буває досить рідко. Як правило, систематизатор ознайомлюється тільки з назвою та анотацією, оскільки прочитати весь документ повністю не вистачає часу та фізичних можливостей. Іноді систематизатор може врахувати бачення автора і присвоїти УДК, на якому наполягає автор. Такі чинники є основними причинами хибного присвоєння індексу УДК.

Системи комп'ютерного аналізу тексту (контенту) позбавлені перелічених недоліків. Тобто комп'ютерна програма з легкістю, за лічені секунди чи навіть долі секунди, може неупереджено обробити текстовий файл будь-якого розміру та виокремити характеристики, які становлять інтерес. Самостійно присвоїти УДК – завдання, яке штучному інтелекту поки що не під силу (хоча досвідчені систематизатори стверджують, що послідовність дій, яку вони виконують, присвоюючи УДК, насправді не така складна, як здається на перший погляд). Тому про заміну систематизатора бібліотеки штучним інтелектом не йтиметься ще довгий час. Створити програмний інструмент, який гармонійно поєднував би можливості неживого комп'ютера та людського творчого потенціалу систематизатора – завдання цілком посильне.

Аналіз досліджень та публікацій. Контент-аналіз або аналіз змісту документа в широкому розумінні – це методика дослідження, аналізу та інтерпретації різних форм людських комунікацій, таких як текстові, графічні, аудіо матеріали. Дана методика охоплює системний аналіз змісту цих матеріалів, виявлення закономірностей, тем, інших відповідних характеристик та формулювання висновків на основі отриманих результатів [6].

У вужчому розумінні, коли розглядаються лише текстові матеріали, контент-аналіз передбачає систематичне, відтворюване стиснення великої кількості слів у меншу кількість категорій змісту документа, що ґрунтується на явних правилах кодування [14].

Фахівці виділяють два основних типи контент-аналізу: кількісний і якісний. Кількісний аналіз націлений на виявлення частоти окремих тем, слів або символів, що містяться у тексті, якісний – фіксує нетривіальні висловлювання, мовні інтонації, цінність змісту повідомлення. Обидва типи аналізу полягають у перетворенні неструктурованого тексту в структуровані дані, або, іншими словами – у знаходженні кількісних характеристик тексту з наступною інтерпретацією отриманих результатів.

Країною, в якій було винайдено контент-аналіз, небезпідставно вважають США, хоча подібний підхід почали використовувати ще в 1640 р. у Швеції, де теологи під час дискусії з офіційним лютеранством порівнювали збірку релігійних гімнів «Пісні Сіону» з гімнами офіційної церкви та підраховували кількість основних релігійних ідей і аналізували їхнє висвітлення: позитивне, негативне чи нейтральне [12].

У подальшому методи контент-аналізу набули розвитку у статтях Дж. Спіда «Чи дають тепер газети новини?» (1893 р.), де було проаналізовано випуски

нью-йоркських газет за 1881 – 1883 рр. [9], і Д. Вілкокса – «Американська газета у світлі соціальної психології» (1900 р.) та ін. [10]. Дж. Спід виміряв обсяг матеріалів за кожною темою у дюймах і порівняв результати. Виявилося, що газети стали приділяти більше уваги пліткам та скандалам, але менше – літературі, політиці, релігії.

В Європі ідеї контент-аналізу стали відомими завдяки Максу Веберу [12], який на першому засіданні Німецького соціологічного товариства у 1910 р. закликав використовувати контент-аналіз для оцінки охоплення пресою політичних акцій у Німеччині та вивчення громадської думки.

Згодом контент-аналіз почали застосовувати для встановлення достовірності історичних документів. Дослідники підраховували частоту вживання слів у документах сумнівного походження і порівнювали з частотою вживання відповідних слів у документах, достовірність яких сумнівів не викликала.

Ідеї контент-аналізу знайшли застосування в бізнесі. Для бізнесу великі обсяги даних, що генеруються кожного дня, створюють одночасно і широкі можливості, і численні проблеми. З одного боку, дані допомагають компаніям сформувати уявлення про думку споживачів щодо їхніх товарів та послуг. З іншого – створюють проблему: – необхідність оперативного оброблення величезного масиву неструктурованої інформації. Подібна ситуація спостерігається в науковій та бібліотечній сферах: з одного боку, в неймовірно великому масиві найрізноманітніших текстових файлів знаходяться і ті, що містять потрібні для користувача інформацію та знання. З іншого – завдання пошуку саме тієї інформації, яка становить інтерес у даний момент, доволі непросте.

Ймовірно, однією з перших комп'ютерних програм, створених для проведення контент-аналізу, слід вважати General Inquirer, яку розробив у 1962 р. Ф. Стоун [7]. Ця програма функціонує і дотепер, вона використовує 182 семантичні категорії та словник обсягом у декілька сотень тисяч слів, які поставлені у відповідність цим категоріям.

Програмних пакетів для контент-аналізу існує велика кількість, мова йде про сотні одиниць. Умовно їх можна розділити на повністю автоматизовані потужні комерційні пакети для контент-аналізу та прості універсальні пакети для підрахунку частот різних слів у текстах. Крім того, розроблено онлайн-системи для контент-аналізу. Осторонь стоять програми для оптимізації сайтів з метою більш високого ранжування в пошукових системах. Йдеться про так звану SEO-аналітику (SEO – Search Engine Optimization).

Повністю автоматизовані пакети для контент-аналізу включають в себе розроблені авторами програм аналітичні словники, в які, зазвичай, неможливо або досить важко внести зміни. При цьому подібні програми не можуть розуміти текст у людському значенні цього слова. Завдання інтерпретації масиву документів – сфера відповідальності аналітика. Прикладами таких суцільно автоматичних «контент-аналітиків» є програми WordStat, Crawdad Text Analysis System, Diction, CATPAC та ВААЛ (одна з небагатьох програм для аналізу російськомовних та україномовних текстів).

Інший тип програм – прості універсальні пакети підрахунку частоти різних слів у текстах. До недоліків таких програм слід віднести відсутність категоризації підрахованих слів, що призводить до викривлення результатів, оскільки враховуються окремо беззмістовні слова, до яких належать службові частини мови. Крім того, різні форми змістовних слів рахуються як різні слова, тому кількість одиниць підрахунку не дорівнює кількості значень. Останній зі згаданих недоліків був частково усунутий розробниками через процедуру лематизації (виділення незмінних частин слів).

Однак подібних програм з лематичним вдосконаленням не існує для слов'янських мов, які мають надзвичайно гнучку морфологію, що передбачає досить відмінні форми для різних родів, чисел, відмінків тощо. Прикладами такого типу програм є Yoshikoder, Concordance, HAMLET. Ці програми – універсальні і не прив'язані до певної мови.

Коротко торкнемося програм, які належать до двох із вказаних типів.

Concordance – програма, яка використовується для проведення контент-аналізу електронних документів [5]. У ній можна створювати списки пов'язаних одиниць підрахунку, індексів, слів при роботі з електронним текстом. Вона дає змогу обробляти великі масиви інформації, уможливлює перегляд кореляцій між словами, що входять у словник контент-аналізу. Результати роботи можуть бути розміщені в Інтернеті за допомогою вбудованих можливостей програми.

Yoshikoder – багатомовна програма контент-аналізу, розроблена для різних програмних платформ [8]. Була створена при реалізації проєкту Identity Project у Гарварді. Програма безкоштовна, дає змогу завантажувати документи, створювати та використовувати контент-словники, вивчати ключові слова в контексті та виконувати контент-аналіз на будь-якій мові. Yoshikoder працює з текстовими документами в звичних ASCII, Unicode (наприклад, UTF-8) кодуваннях або з національними таблицями кодування, наприклад, Big5 Chinese. Рідним форматом файлів Yoshikoder є формат XML, тому словники і файли ключових слів не є пропрієтарними, вони зручні для читання. Результатом роботи Yoshikoder є резюме документа у вигляді таблиці частотності слів або згідно заданого словника для контент-аналізу.

Компанія IBM (https://www.ibm.com/) розробила систему IBM Watson – бізнес-аналітика, пошук даних і аналіз тексту [11]. Вона дає змогу виявляти глибші відомості, приховані у ділових документах, з меншими зусиллями, підвищує продуктивність працівників завдяки автоматизованому пошуку інформації та розуміння за допомогою оброблення природної мови з використанням штучного інтелекту на базі користувацьких моделей NLP і великих мовних моделей (LLM) від IBM Research.

Свого часу використання комп'ютерної техніки поширилось з бізнесу та банківської сфери на роботу наукових і бібліотечних установ. Так і тепер методи контент-аналізу, первинно розвинені для задоволення інформаційних потреб бізнесу, поступово охоплюють наукову та бібліотечну справу [2]. Вищезгадані системи не пристосовані для вирішення проблем, що постають перед бібліотеками,

користувач обмежений можливостями системи та вимогами до формату вхідних текстових файлів (наявність структурованих бібліографічних даних тощо).

Мета статті – запропонувати методику аналізу достовірності індексу УДК повнотекстових документів, що надходять до бібліотеки від різних організацій і авторів в електронному форматі, продемонструвати її застосування на прикладі п'яти електронних документів економічної тематики (індекс УДК 331), використовуючи створений програмний інструмент «Аналіз текстів».

Виклад основного матеріалу. При аналізі текстових даних фахівці користуються поняттям косинусної подібності – міри подібності між двома ненульовими векторами. Косинусною мірою, або мірою подібності числових векторів, називають їх скалярний добуток, поділений на добуток їх модулів. Іншими словами, для двох заданих векторів A і B косинус подібності, cos(θ), може бути представлений через скалярний добуток та довжину:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

де Ai та Bi координати вектора A і B відповідно [4].

Косинусна подібність не залежить від величини векторів, а лише від кута між ними. Косинус кута може набувати значення від -1 (включно) до 1 (включно). Косинусна подібність двох пропорційних векторів становить 1, двох перпендикулярних векторів – 0 та -1 для двох протилежних векторів. У контексті аналізу текстових даних значення координат векторів не можуть бути від'ємними, тому косинусна подібність обмежена замкненим інтервалом від 0 до 1.

Добуток та інші математичні операції для складових частин текстів, тобто для речень чи слів, безпосередньо не означені. Отже, з текстовими файлами потрібно якимось чином співставити відповідні числові вектори. Такий процес носить назву векторизації. Кожне слово розглядається як відповідна координата, з кожним словом зіставляється відповідна частота його вживання в тексті, тобто текстовий документ представляється вектором кількості вживань кожного слова в документі. Кількість або частота вживань слова є числом більше нуля, саме тому міра косинусної подібності може набувати значення від нуля до одиниці. Таким чином, поняття косинусної подібності відображає міру подібності двох документів з точки зору їх тематики і не залежить від довжини текстових документів.

Знаходження косинусної міри подібності текстів передбачає певні кроки. Насамперед ретельно підбирається сукупність текстів, які якнайточніше відповідають темі, що досліджується. Таку сукупність називають корпусом [1]. Наступним кроком текст корпусу перетворюється в словник. Слід зауважити, що в корпусі слова, які мають однакове значення, будуть вживатися в різних формах (відмінках та дієвідмінах, числах, родах). На цьому етапі текст нормалізується, тобто вся сукупність існуючих форм або лексем замінюється одним словом. Крім того, при векторизації тексту слід видалити малоінформативні слова (далі – стоп-слова), наприклад, прийменники, займенники, прислівники, сполучники та знаки пунк-

туації. У результаті порядок слів у тексті буде порушено, залишиться так званий «мішок слів» – невпорядкована сукупність слів. І от тепер можна отримати кількісні характеристики – порахувати кількість вживань кожного слова в тексті.

Такі ж перетворення слід провести у тексті, який аналізується. Далі кожне слово інтерпретується як розмірність вектора, якщо кількість унікальних слів після векторизації становить, скажімо, m. Це означає, що розглядається вектор у m-мірному просторі. Кількість вживань кожного слова інтерпретується як відповідна координата вектора у m-мірному просторі.

Стосовно тексту, що аналізується, вирахувані значення косинусної подібності можуть бути інтерпретовані таким чином: 0 – файл, що аналізується, не має жодного спільного слова з корпусом; 1 – слова корпусу і тексту, що аналізується, а також частота їх вживання повністю однакові, тобто має місце випадок плагіату.

Стосовно текстів, що аналізуються, далі вживатиметься термін КТН – коефіцієнт тематичного напряму – косинусна міра подібності, як характеристика подібності текстових векторів.

Програма «Аналіз текстів» (авторська розробка) – призначена для формування частотних масивів текстових файлів, які використовуються для контент-аналізу. Інтерфейс програми представлено на рис. 1.

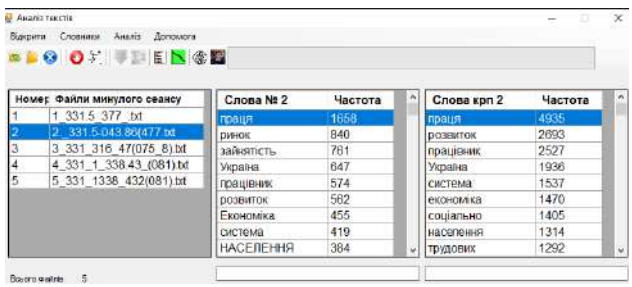

Рис. 1. Інтерфейс програми «Аналіз текстів»

Кнопка «Лексеми»  – робота зі словником лексем здійснюється за допомогою окремого модуля «Лексеми», який входить у «Аналіз текстів» (див. рис. 2).

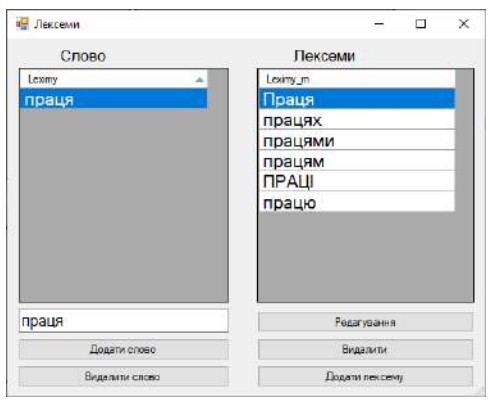

Рис. 2. Модуль «Лексеми»

За допомогою словника лексем у частотних масивах відбувається заміна всіх лексем (словоформ) на слово синонім та перераховується частотний масив. При контент-аналізі на плагіат або новизну статті ці корегування не залишають шансів так званим копіювальникам, навіть з використанням штучного інтелекту. Тут у авторів – тільки один вихід: включати свій мозок і створювати новизну.

Кнопка «Стоп-слова» – словник стоп-слів, оперується за допомогою окремого модуля «Стоп-слова», який також входить у «Аналіз текстів» (див. рис. 3). За допомогою словника стоп-слів у частотних масивах відбувається видалення зайвих слів.

Рис. 3. Модуль «Стоп-слова»

Програма «Аналіз текстів» може застосовуватись на будь-яких тематичних напрямах зі своїми специфічними словниками лексем. Всі дані зберігаються в базі даних SQLite – один файл, який фізично розташовано у папці програми. Кожна копія програми «Аналіз текстів» працює зі своїм набором файлів, що аналізуються, для яких створюються унікальні словники стоп-слів та лексем.

Для аналізу візьмемо, наприклад, файли з повнотекстової бази даних із УДК 331 (Праця. Наука про працю. Економіка праці. Організація праці):

1. Інститут економіки та прогнозування НАН України. Ринок праці та професійно-технічна освіта: особливості взаємодії [Електронний ресурс] : монографія / ДУ «Ін-т екон. та прогнозув. НАН України». Київ, 2021. 297 с. ISBN 978-966-02-9806-4;

2. Інститут економіки та прогнозування НАН України. Інституційні засади розвитку ринку праці в Україні [Електронний ресурс] : монографія / ДУ «Ін-т екон. та прогнозув. НАН України»;

3. Економіка праці й соціально-трудові відносини [Електронний ресурс] : навчальний посібник / В. А. Ткачук, Є. О. Ланченко, О. Д. Балан, І. П. Гаврилюк. Київ : НУБіП України, 2022. 616 с.: табл., рис. Бібліогр.: с. 531–536 (77 назв). ISBN 978-617-8102-17-3;

4. Ланченко Є. О. Розвиток соціально-трудових відносин в аграрній сфері [Електронний ресурс] : монографія. Київ, 2022. 266 с. ISBN 978-617-8007-53-9.

5. Ланченко Є. О. Формування системи соціально-трудових відносин у аграрному секторі економіки [Електронний ресурс] : монографія. Київ, 2019. 556 с. ISBN 978-966-929-909-3;

Результати основних показників занесемо у табл. 1.

Таблиця 1

Основні показники обробки текстів

№	Назва	КТН	УДК
1	Ринок праці та професійно-технічна освіта: особливості взаємодії	0,472	331.5:377
2	Інституційні засади розвитку ринку праці в Україні	0,863	330.341.42: 331.5-043.86(477)
3	Економіка праці й соціально-трудові відносини	0,837	331:316.47(075.8)
4	Ланченко Є. О. Формування системи соціально-трудових відносин у аграрному секторі економіки	0,853	331.1:338.43(081)
5	Ланченко Є. О. Розвиток соціально-трудових відносин в аграрній сфері	0,831	331.1:338.432(081)

КТН кожного текстового файлу визначається відносно свого корпусу, текстовий вектор якого формується з усіх файлів крім того, що аналізується.

Кнопка – графік КТН – розрахунок КТН всіх файлів (частотних масивів).

Кожен КТН розраховується для двох векторів: частотний масив файлу та частотний масив корпусу – сумарний частотний масив, створений з інших файлів. Дані таблиці 2 представлені на рис. 4.

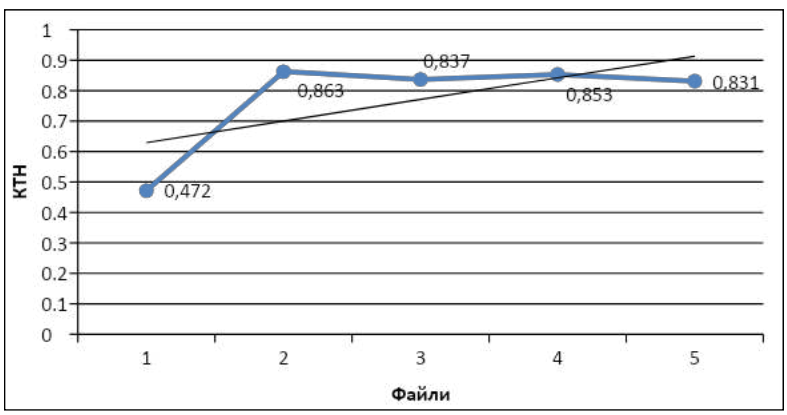

Рис. 4. Розподіл КТН файлів

Лінійну апроксимацію розподілу КТН: **y = 0,0709x + 0,5585** було отримано в Excel. Як видно з рис. 4, КТН файлу №1 відрізняється від загального тематичного напряму, таким чином значно збільшує коефіцієнт при **x**. Щоб визначити оптимальний коефіцієнт при **x,** проведемо аналіз вищезгаданих файлів, але без файлу №1. В ідеальному випадку цей коефіцієнт при **x** дорівнював нулю – горизонтальна лінія апроксимації. Результати перерахунку нової групи файлів програмою «Аналіз текстів» наведено на рис. 5.

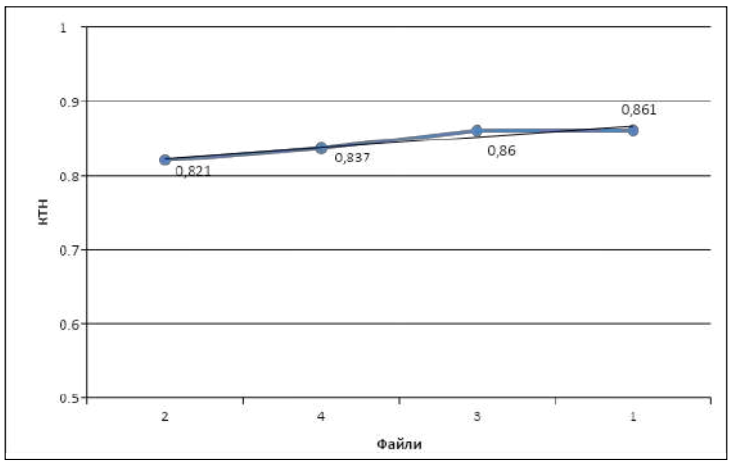

Рис. 5. Розподіл КТН файлів без файлу №1

Як видно з рис. 5, коефіцієнт при **x** дорівнює 0,0143, а з файлом №1 – 0,0709. 0,0709/0,0143 = 4,958, тобто більший майже у п’ять разів. Робимо практичний висновок: для одного тематичного напряму достатньо, щоб коефіцієнт при **x** лінійної апроксимації був 0,015, який гарантує: оброблені файли (повнотекстові надходження) були одного тематичного напряму. В результаті, дослідження отримали об’єктивний критерій кластеризації текстів.

Розглянемо частотні масиви слів файлів за номерами 2, 3, 4, 5.

Таблиця 2

Фрагмент частотного словника файлу: №2; КТН = 0,863; УДК 331.5-043.86(477)

Слово	Частота
праця	1720
працівник	1005
ринок	840
зайнятість	761
Україна	647
трудових	614
розвиток	562
сфера	464
оплата	426
система	419
економіка	373
освіта	305

Таблиця 3

Фрагмент частотного словника файлу: №3; КТН = 0,834; УДК 331:316.47(075.8)

Слово	Частота
праця	3048
працівник	2268
робота	1127
трудових	1082
оплата	970
підприємства	701
СОЦІАЛЬНО	647
Україна	637
система	551
відносин	544
часу	531
розвиток	515
виробництва	490

Таблиця 4

Фрагмент частотного словника файлу: №4; КТН = 0,853; УДК 331.1:338.43(081)

Слово	Частота
розвиток	671
працівник	576
сільського	556
праця	510
сільськогосподарське	336
СТВ	336
Україна	328
населення	309
трудових	288
підприємства	283

Таблиця 5

Фрагмент частотного словника файлу: №5; КТН = 0,831; УДК 331.1:338.432(081)

Слово	Частота
працівник	1371
розвиток	1259
праця	1191
сільського	1155
підприємства	772
СТВ	762
Україна	714
сільськогосподарське	693
економіка	677
трудових	676

Таблиця 6

Фрагмент частотного словника файлу №1; КТН = 0,472; УДК 331.5:377

Слово	Частота		Слово	Частота
освіта	1010		працівник	231
навчання	667		учень	231
професійної	472		ринок	223
ПРОФЕСІЙНО	404		підготовки	183
кваліфікація	383		зварювання	170
технічна	381		сфера	164
заклад	340		забезпечення	155
праця	334		діяльності	151
професія	318		області	147
робота	286		професійних	131
система	257		навчально	131
Україна	257		процесу	115
розвиток	248		навички	110
			професійного	110

Як бачимо, контент файлу №1 «Ринок праці та професійно-технічна освіта: особливості взаємодії» не відповідає заданому тематичному напряму, тобто, його УДК слід відкорегувати (скорегований УДК 377:331.5). Проаналізувавши частотний масив даної роботи, можемо стверджувати, що автор більше уваги приділяв освіті, а не ринку праці. Систематизатор не має фізичних можливостей читати кожну роботу повністю, тому визначив індекс УДК з назви та анотації або зі слів автора, чого явно недостатньо для достовірного присвоєння УДК.

Цінність (важливість) точності систематизації зростає зі зростанням кількості повнотекстових одиниць і є вдалим прикладом співпраці комп'ютера і людини.

При кожному певному науковому дослідженні як бонус можна отримати частотні масиви слів, які відповідають конкретному УДК, а також тематичні масиви стоп-слів і лексем. Їх можна визначити як бібліотечну кластеризацію, яка не лише дає систематизатору інструмент для визначення УДК, бере на себе рутинні процеси, але й допомагає визначити кількість кластерів для будь-якої електронної бібліотеки. За результатами комп'ютерного аналізу автоматично визначається кількість кластерів (предметних рубрик), які раніше визначав систематизатор на власний розсуд.

Частотні масиви слів певних тематичних напрямів також можуть бути застосовані при комплектуванні повнотекстових баз даних з певним критерієм відбору літературних одиниць. Для конкретного надходження створюватиметься частотний масив, КТН якого буде розраховано з відповідним корпусом. Залежно від отриманого КТН комплектатор ухвалить відповідне рішення щодо включення електронного видання до бази даних.

Висновки. Запропоновану методику корекції УДК можна застосовувати у разі створення репозитаріїв електронних текстів. Це сприятиме підвищенню якості інформаційного пошуку та вибору контенту. Кожний наступний електронний документ доцільно аналізувати на відповідність УДК, присвоєного систематизатором, який визначався за назвою та анотацією, з УДК відповідного корпусу.

При накопиченні певної кількості електронних документів, завдяки розробленій методиці, УДК нового тексту (надходження) можна визначити автоматично за показником коефіцієнтів тематичного напряму (близько одиниці) нового тексту та відповідного корпусу.

Вектор КТН текстів, що досліджувались, їх розподіл за зростанням коефіцієнтів тематичного напряму, дав змогу виявити кластер – групу однакових текстів за змістом. Достовірним критерієм є величина коефіцієнту при змінній лінійної апроксимації, в ідеалі горизонтальна поличка на графіку розподілу КТН – коефіцієнт дорівнює одиниці. Кількість тематичних напрямів визначається кількістю кластерів.

Список бібліографічних посилань

1. Дані текстових корпусів у лінгвістичних дослідженнях: монографія / В. А. Широков, І. В. Шевченко, А. П. Загнітко та ін.; Національний ун-т «Львівська політехніка». Львів : Вид-во Львів. політехніки, 2015. 160 с.

2. Симоненко Т. В. Мережеве інформаційно-бібліотечне забезпечення наукових досліджень: автореф. дис. канд. наук із соц. комунікацій / НАН України, Нац. б-ка України ім. В. І. Вернадського. Київ, 2011. 18 с.

3. Chatbot GPT. URL: https://chatgpt.org.ua/.

4. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. 506 p.

5. Concordance. URL: https://www.concordancesoftware.co.uk/.

6. Content Analysis – Methods, Types and Examples. URL: https://researchmethod.net/content-analysis/.

7. Descriptions of Inquirer Categories and Use of Inquirer Dictionaries. URL: https://inquirer.sites.fas.harvard.edu/homecat.htm.

8. Lowe W. Yoshikoder: Cross-platform multilingual content analysis'. Java software version 0.6.5. 2015. URL: https://yoshikoder.org.

9. News about News. URL: https://www.tandfonline.com/doi/abs/10.1080/00947679.2001.12062572.

10. The American Newspaper: a Study in Social Psychology. URL: https://journals.sagepub.com/doi/10.1177/000271620001600104.

11. Watson. URL: https://www.ibm.com/watson.

12. Weber R. P. Basic Content Analysis. Beverly Hills, CA: SAGE. 1990. 96 p.

13. Worldwide IDC Global DataSphere Forecast, 2022–2026. URL: https://www.idc.com/getdoc.jsp?containerId=US49018922.

14. Zaidman-Zait A. Content Analysis. 2014. URL: https://doi.org/10.1007/978-94-007-0753-5_552.

References

1. Shyrokov, V. A., Shevchenko, I. V. & Zahnitko, A. P. (2015). Dani tekstovykh korpusiv u linhvistychnykh doslidzhenniakh: monohrafiia [Text corpora data in linguistic research: monograph]. Lviv, Ukraine: Vyd-vo Lviv. politekhniky. [In Ukrainian].

2. Symonenko, T. V. (2011). Merezheve informatsiino-bibliotechne zabezpechennia naukovykh doslidzhen [Network information and library support for scientific research]. (Extended abstract of PhD disertation). V. I. Vernadskyi National Library of Ukraine. Kyiv, Ukraine. [In Ukrainian].

3. Chatbot GPT. Retrieved from https://chatgpt.org.ua [In English].

4. Manning, C., Raghavan, P., & Schütze. H. (2008). Introduction to Information Retrieval. Cambridge University Press. [In English].

5. Concordance. Retrieved from https://www.concordancesoftware.co.uk [In English].

6. Content Analysis-Methods, Types and Examples. Retrieved from https://researchmethod.net/content-analysis [In English].

7. Descriptions of Inquirer Categories and Use of Inquirer Dictionaries. Retrieved from https://inquirer.sites.fas.harvard.edu/homecat.htm [In English].

8. Lowe, W. (2015). Yoshikoder: Cross-platform multilingual content analysis. Java software version 0.6.5. Retrieved from https://yoshikoder.org. [In English].

9. News about News. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/00947679.2001.12062572 [In English].

10. The American Newspaper: a Study in Social Psychology. Retrieved from https://journals.sagepub.com/doi/10.1177/000271620001600104 [In English].

11. Watson. Retrieved from https://www.ibm.com/watson [In English].

12. Weber, R. P. (1990). Basic Content Analysis. Beverly Hills, CA: SAGE. [In English].

13. Worldwide IDC Global DataSphere Forecast, 2022–2026. Retrieved from https://www.idc.com/getdoc.jsp?containerId=US49018922 [In English].

14. Zaidman-Zait, A. (2014). Content Analysis. Retrieved from https://doi.org/10.1007/978-94-007-0753-5_552 [In English].

Oleksandr Kuznetsov,
https://orcid.org/0000-0002-9902-1295,
Head of Department of Scientific Document Processing and Cataloging,
State Scientific and Technical Library of Ukraine (Kyiv, Ukraine)

Victor Zaika,
https://orcid.org/0009-0003-6582-6524,
Candidate of Physical and Mathematical Sciences,
Engineer of the Department of Bibliometrics and Scientometrics,
V. I. Vernadskyi National Library of Ukraine (Kyiv, Ukraine)

UDC CODE DETERMINATION OF NEW ELECTRONIC RECEIPTS FOR THE FORMATION OF AN ELECTRONIC LIBRARY BY MEANS OF SOFTWARE

The purpose of the article is to propose a validation technique of the UDC index of library electronic documents accessions and to demonstrate its usage for the five electronic documents on economic topics (UDC index 331) based on the developed software tool «Text Analysis». **Research methodology.** The quantitative method of document content research is applied. To find documents (files) similar in content, the concept of the cosine measure of similarity was used and coefficients of the thematic direction, were calculated for each document. Text files were vectorized, that is, represented as vectors in a multidimensional space. For this purpose, different word forms were reduced to one lexeme and the number (or frequency) of lexeme usage in each document was calculated. Lexemes are interpreted as coordinates, and the frequency of use is interpreted as the value of the corresponding coordinate. After vectorization of the texts, the mathematical apparatus of analytical geometry was applied, and a numerical value - the coefficient of the thematic direction - was matched to the topic of each text document. **Scientific novelty.** For the first time, methods of content analysis, namely, quantitative analysis, were used to assess the reliability of the UDC index of a document, and a software tool was created, the use of which will help the systematizer to confirm or refute the UDC index of a dubious document without reading it. **Conclusions.** The author's software tool and the proposed UDC correction technique can be used when creating repositories of electronic texts and will contribute to improving the quality of information search and content selection. When accumulating a certain number of electronic documents, thanks to the developed methodology, the UDC of a new text (receipt) can be determined automatically by the indicator of the coefficients of the thematic direction (close to one) of the new text and the corresponding corpus. The vector of coefficients of the thematic direction of the studied texts, their distribution according to the growth of the coefficients of the thematic direction, made it possible to identify a cluster - a group of texts with the same content. A reliable criterion is the value of the coefficient for a variable linear approximation, ideally a horizontal shelf on the graph of the distribution of the coefficients of the thematic direction - the coefficient is equal to one. The number of thematic areas is determined by the number of clusters.

K e y w o r d s : computer text analysis systems, content analysis, cosine similarity measure, UDC index, cluster, electronic library, frequency array, coefficient of thematic direction, software packages for content analysis.

State Scientific and Technical Library of Ukraine
180, Antonovycha str., Kyiv, 03150, Ukraine
e-mail: nkof@dntb.gov.ua

V. I. Vernadskyi National Library of Ukraine
3, Holosiivskyi Avenue, Kyiv, 03039, Ukraine
e-mail: victor.zayika@gmail.com

Стаття надійшла до редакції 18.07.2023 р.